DOCUMENT RESUME

ED 208 029 TM 810 724

AUTHOR Mead, Ronald J.
TITLE Basic Ideas in Item Banking.
SPONS AGENCY National Board of Medical Examiners, Philadelphia, Pa.; National Inst. of Education (ED), Washington, D.C.
PUB DATE 14 Apr 81
GRANT NIE-G-89-0078
NOTE 19p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Los Angeles, CA, April 14-16, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Equated Scores; Goodness of Fit; *Item Banks; Latent Trait Theory; Test Construction; Test Items; *Test Validity
IDENTIFIERS *Calibration; *Rasch Model

ABSTRACT

The central idea in building and maintaining an item bank is to calibrate all the items onto a "common variable." The arithmetic involved in the calibration process is presented. It is recommended that an analysis of fit be done in every application to verify that the estimates of item difficulties are in fact sample-free. These procedures are explained. Once an item bank is built, a common calibration for all items should be established and routinely checked. Special procedures for adding new items, updating old items, and dropping obsolete items are described. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED208029

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

BASIC IDEAS IN ITEM BANKING

Ronald J. Mead

MESA Psychometric Laboratory
Department of Education
University of Chicago

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. J. Mead

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 724

The preparation of the document was supported by
National Institute of Education Grant NIE-G-89-0078
and the National Board of Medical Examiners

Paper presented at the NCME Annual Meeting
Los Angeles, California
April 14, 1981

BASIC IDEAS IN ITEM BANKING

Ronald Mead
MESA Psychometric Laboratory
Department of Education

University of Chicago

The central idea in building and maintaining an item bank is to "calibrate" all the items onto a "common variable". The arithmetic involved in the calibration process is well known and straightforward (Choppin, 1968; Wright and Stone, 1979; Rentz and Bashaw, 1977; Mead and Kreines, 1978) so I will deal with that first. The implication in the phrase "common variable" is the notion that all the items measure the same thing. Establishing that this is reasonable goes beyond calibration and is normally called something like "item fit analysis" but "validation" might be a more appropriate name. I will consider that later.

CALIBRATION

Calibrating a Single Form

When all the people take all the items, you have an "item bank" as soon as you have computed estimates of the difficulties. There are a number of ways that this can be accomplished by hand or by one of several computer programs (eg., BICAL). In the process, the origin of the scale is set at the average item difficulty but this is only a numeric convenience.

The number associated with an item is the distance from the center of the form to the item in question. A negative value in-

dicates that the item is easier than the average and a positive value indicates that the item is harder than the average.

Calibrating Two Forms

If the items are in two forms instead of just one, the first step is the same: calibrate each form separately. We then have two banks, each with its mean difficulty set to zero. Combining them requires finding the distance between these two origins. For this to be possible, the two calibrations must have something in common. This can be either common items or common persons.

To illustrate the idea, consider two forms, A and B, shown in Figure 1, with a single common item linking them. To give it a name, let's call it 'Item 7' and assume it has an estimated difficulty of +1.0 in Form A and -0.5 in Form B. In other words, the distance from the center of Form A to Item 7 is 1 logit and the distance from Item 7 to the center of Form B is another half logit. This makes the distance from the center or "origin" of Form A to the origin of Form B:

$$1.0-(-0.5) = 1.5 \text{ logits.}$$

The only sleight of hand is that I was careful to change the sign of the half logit to show that I was going from the item to the origin of Form B rather than from the origin to the item.

If there are several common items, then I would work with their average difficulty but the logic is unchanged. Similarly if

there is a group of common people rather than common items, I would work with their average ability. The basic process is the same in any case: Find the distance from the first origin to the point in common and then the distance from the common point to the second origin.

The sum of these distances is typically referred to as the "link" between the two forms (sometimes it is called the "translation constant" or the "shift"). The way I have arranged it here, it is the amount that should be added to the difficulties of all the items in Form B to shift them onto the origin of Form A. (There is nothing sacred about that particular origin, however; we can shift it to some more convenient point if there is any reason to do this.)

The complication remaining is what to do with the pair of difficulties we now have for each of the common items. Because these difficulties were estimated from different data, they will never be exactly the same. Unless there is some reason to prefer one calibration over the other, a reasonable thing to do is to take a weighted average using the inverse square of the standard errors of calibration as the weights. This weighs each estimate by the amount of information it contains and takes account of both how large and how relevant each sample is. The inverse square root of the sum of weights is then the standard error for the pooled estimate.

Calibrating Several Forms

Establishing an item bank usually requires more items than can be given in one or two forms. When several forms are involved, we begin in the same way:

a) Calibrate each form separately, and

b) Find the link between each pair of forms that have a common point,

Then, because we are dealing with data, the set of links will not be consistent. For example, in Figure 2, linking Form A to Form B, then Form B to Form C, and finally Form C to Form A amounts to linking Form A to itself and so the sum of those links should be zero. However, it can never be exactly zero, so we need a procedure to resolve the inevitable inconsistencies.

Engelhard and Osberg (1981) give the general least squares answer, but a procedure (Wright and Stone, 1979), which gives the same result and avoids matrix algebra, is:

1) Construct the matrix of link constants t(i,j) (the distance to Form i from Form j).

2) Fill in a good guess for any link that is missing. (Use zero, if you have no better idea.)

3) Compute row means T(i) for the entire matrix exactly as though it were full. (Include the diagonal which is always zero.)

4) Fill in estimates for the missing links computed as

$$t(i,j) = T(i) - T(j)$$

5) Repeat Steps 3 and 4 until the matrix stabilizes.

6) Translate difficulties on Form i to the center of the bank by adding the mean for row i.

$$d(i) = d(i) + T(i)$$

Figure 2 and Table 1 illustrate this for a network of five forms. Some attention must always be given to the direction of the arrows and the signs of the numbers.

In Table 1, I started with zeros for the missing links. It then took ten steps to stabilize. It would have been more reasonable and quicker to make some intelligent guesses from Figure 2, say, -2.0 for link AE; -1.4 for link AD and -1.6 for link BE. Exactly the same thing could also have been done from Table 1 by, for example, subtracting CE from CA to obtain an estimate of AE.

The row mean for Row i is the number that should be added to the difficulties of all items on Form i to shift them onto the common origin, which in this procedure is the center of all the forms. There is no magic in this particular origin. Once we have established a common origin we can shift it anywhere that is convenient for our purposes.

ANALYSIS OF FIT

So far this has been nothing more than elementary arithmetic and good housekeeping. The only hard part is keeping the signs

straight. However, it has all been based on the proposition that the data comform to the dichotomous Rasch model. If the items work in a way that is a reasonable approximation to this proposition, then the estimates of item difficulties are in fact sample-free and everything we have done is valid and just as easy as it seems. Otherwise things can become more complicated. We cannot, however, assume that everything is the way we would like it; this must be verified in every application.

There are three phases in the fit analysis but the idea is the same in each of them: Specific objectivity explicitly refers to the freedom from the ability distribution but it also means, for any appropriate sample, freedom from age, grade, school, race, or sex as well. The fit analysis asks if this appears to be the case for the data in hand.

Phase I: Within Form Fit

The first point at which the fit analysis must be done is when calibrating each form. Ideally this would involve checking that the difficulties are invariant with respect to every possible subdivision of the sample. This could be done physically by dividing up the sample into groups defined by ability, race, sex, age, grade, etc., and reestimating the item difficulties within each group. Likelihood ratios could then be formed to test the equality of different sets of difficulties (Gustafsson, 1978) or they can be plotted against each other (Rasch, 1960; Wright, 1968; Wright and Stone, 1979).

A useful shortcut is to use the "Between Score Group" and "Total" fit statistics (as computed by BICAL). Both of these statistics are based on approximate chi-squares derived from the unconditional maximum likelihood equations and are easy to compute.

The between score group analysis automatically divides the sample by ability and explicitly asks if the empirically obtained item characteristic curve approximates the required shape. If it does, then the ability groups agree on the difficulty and we can be as confident of our estimate as their standards error permit.

The total fit statistic is an attempt to cover everything else without being explicit about it. While it is not particularly sensitive to subtle departures from objectivity, it detects irregularities which threaten the basic meaning of the data quite well.

Formal tests of significance are not of much interest here for three reasons. It is not clear what the null distributions of the individual item statistics really are. Even if it were, null distributions would not help much since we really want to do a series of sequential tests on the items. And we can always make any amount of irregularity acquire any amount of significance by adjusting the sample size.

I depend much more on examining plots. Rather than arbitrarily excluding every item we think is more than two or three standard errors away from expectation, we can plot the various fit statistics against each other and look for points in the plot that are outliers from whatever the distribution is. I am not too concerned if the distribution is fatter than it is supposed to be as long as it seems to be one distribution. The distribution being overweight does influence my opinion of the standard errors, however.

Items identified as misfitting in this manner are almost always easily diagnosed if we are willing to look hard enough. They are items that are miskeyed, that have no right answer, that have more than one right answer, that have a smart way to find the wrong answer, or that have an interaction with special instruction or experience. Recognising the items require investigation from histograms of the fit statistics is straightforward. Correcting or eliminating them can be done in comfort when we have discovered the particular events that produced their aberent performance. Going beyond this and successively rejecting each "next worst fitting" item becomes both statistically and substantively uncertain with no clear stopping points.

Phase II: Within Link Fit

Once we have satisfied ourselves that the items calibrated

for each form are sufficiently consistent, we can begin linking the ones with common points. For each replication within the connecting elements, we have another level of fit analysis. For common items, we are asking whether the two samples (the one that took Form A and the one that took Form B) define the same scale. This is a form of between group fit analysis where the groups are defined by occasion.

This is most easily investigated in a picture of the link made by plotting the two sets of difficulties against each other. The points in this plot should follow (within standard errors) a straight line with slope one and intercepts t(i,j) and t(j,i). Items which stand away from this line do not have "occasion-free" calibrations.

The analysis of fit can be done in a manner analogous to that described for within form. Rather than imposing an absolute standard, look for items that are obviously different than the others without worrying too much about where the standard error control lines actually fall. Items identified by this approach are usually easy to explain.

Items which do not fit in a link usually turn out to be:

i) different items that were given the same name,

ii) items that were printed differently in one form,

iii) items whose answers changed between administrations, or

iv) items that interact with special experience or instruction.

The last category deserves some discussion. Ideally, we would like persons receiving instruction to move forward on the variable but we would not like them to disturb our operational definition of it. If this were really true, then it would not matter when during instruction the items were calibrated. The items, however, are only imperfect instances of the variable, being told the answer to one of them, or even being told how to solve a special class of them, does not necessarily make a person better able to deal with every other item.

For example, in a bank of mathematics items recently constructed at the MESA Psychometric Laboratory (Wright and Stone, 1980), it was found that fraction problems written horizontally were harder than the same problems written vertically for fifth graders, but not for sixth graders. An extraneous variable of practice or familiarity distinguished the two grades with respect to horizontal items. For the fifth graders, they had one difficulty determined jointly by the complexity of the arithmetic and the unfamilarity of the format. For the sixth graders, they had another (lower) difficulty because the format was no longer a factor.

In this case, the items were included in the bank with the sixth grade difficulties. This means that when they are given to

a child who has trouble with the horizontal notation, this will show up in the fit analysis for that child as a cluster of unexpected failures, diagnositic of this child's particular deficiency. For fifth graders, this should not be alarming; for older children, it might warrant some action.

Phase III. Between Link Fit

When we are dealing with a matrix of links (Table 1), we can take the analysis one step further. Since each entry in the matrix can be predicted from the margin, i.e., $t(i,j)=T(i)-T(j)$, we can compute residuals for each of the observed links. The matrix of residuals can then be summarized in whatever manner interests us to check if particular forms, levels or samples seem to present unusual problems (e.g., plot observed against expected)

While I am unable to provide any fool proof rules for detecting misfitting items or links, I cannot over emphasize the importance of performing analyses of misfit. When dealing with a single, fixed form, it is possible to live with a very loose approximation to specific objectivity. However, as item banks grow larger and cover wider ranges, even minor departures from objectivity become important. If you are planning to do "test-freed" measurement, you need a bank well enough constructed to support it. Also the careful investigation of items you are interested in is always instructive. It invariably leads to new insights into the variable and how people relate to it.

Bank Maintainance

The basic ideas in maintaining a bank are the same as for building it. We need to establish a common calibration for all items and we need to check routinely that things are working the way we want. There are a few new details we should think about explicitly.

Adding New Items

New items can be added anytime we like. We need only administer them with some previously calibrated items and use these as a common point with the bank. This amounts to treating the bank as through it were a form which has some items in common with the new form. A link can then be calculated, added to the difficulties of the new items, and the new records inserted into the file. Of course, an analysis of fit will be performed on the old items to assure ourselves that everything is still under control.

Updating Old Items

We again have the problem of what to do with the old items now that we have still another estimate of their difficulties. There are two schools of thought. We can average in the new information using the same weighted average as before. Or we can

leave the bank difficulties as they were. I favor the second approach.

Averaging is appropriate only if we have evidence that the difficulties have not changed (i.e., the test of fit was acceptable). In other words, averaging is appropriate only when it is not necessary (unless of course, we want to decrease the standard error).

Averaging will also create some fuzziness about how to interpret results. A given score on a fixed form will not be associated with exactly the same ability as it was last year. This will be hard to explain to people who are trying to use the results.

Continuous updating of the banked difficulties can have a more dangerous aspect. It can obscure small but real drifts in the difficulties of some items. If there is a slow but systematic change, allowing ourselves to adjust for it automatically may keep us from noticing it.

A more appealing procedure (once we have acceptable standard errors) is to leave the difficulties where they were in the original calibration until we have strong evidence that they have changed. When that happens, we can either drop the item or substitute the new difficulty.

Dropping Obsolete Items

Once an item has become obsolete, it should be eliminated from the bank. The question is not what to do but when. "Obsolete" means that it no longer belongs on the variable we are measuring. This will be the result of the item failing some phase of the fit analysis after it has been reused.

The decision of when to update a difficulty or when to drop an item is rarely obvious. There should be a periodic analysis of each item's behavior over all its administrations. This is a between-occasion analysis and requires only that we save the item's history. When there appear to be differences in the difficulties, then some action is needed. Whether that action is dropping the item or repairing our opinion of it will depend on what we think has happened. This is a substantive question that should be manageable once the statistical analysis has attracted our attention to the item.

References

Choppin, G. An item bank using sample-free calibration, Nature, 219, 1968.

Engelhard, G. and Osberg, D. Constructing a test network with a Rasch measurement model. Paper presented at the Eastern Educational Research Association annual meeting. Philadelphia, 1981.

Gustafsson, J. Testing and obtaining fit of data to the Rasch model. Paper presented at the American Educational Research Association annual meeting. San Francisco, 1978.

Mead, R. and Kreines, D. Linking tests with the Rasch Model. Paper presented at the American Educational Research Association Annual meeting. San Francisco, 1978.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danmarks Paedagogiske Institut, 1960. (Reprinted Univesity of Chicago Press, 1980).

Rentz, R. and Bashaw, W. The national reference scale for reading: An application of the Rasch model. Journal of Educational Measurement, 14, 161-180, 1980.

Wright, B. Sample-free test calibration and test-free measurement. Proceeding of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968.

Wright, B. and Stone, M. Best Test Design. Chicago: MESA Press, 1979.

Wright, B. and Stone, M. The NUMERALS Math Item Bank. Informal report, MESA Psychometric Laboratory, Chicago, 1980.

Figure 1: Linking Two Forms

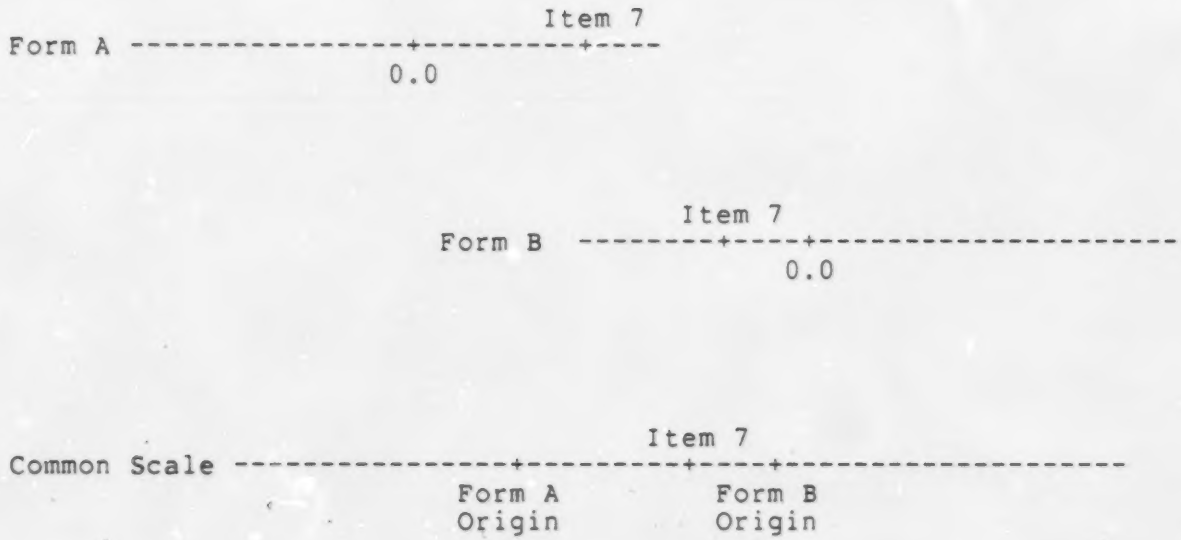

Figure 2: Link Network for Five Forms

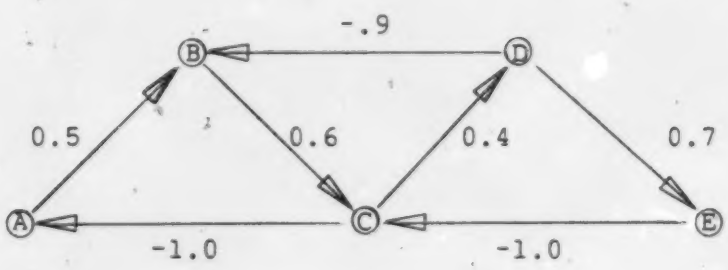

Table 1: Calculating Links for Several Interconnected Forms

Step 0:	A	B	C	D	E	T(I)
A	0	-.5	-1.			-.30
B	0.5	0	-.6	-.9		-.20
C	1.0	0.6	0	-.4	-1.	0.04
D		0.9	0.4	0	-.7	0.12
E			1.0	0.7	0	0.34

Step 1:	A	B	C	D	E	T(I)
A	0	-.5	-1.	-.42	-.64	-.51
B	0.5	0	-.6	-.9	-.54	-.31
C	1.0	0.6	0	-.4	-1.	0.04
D	0.42	0.9	0.4	0	-.7	0.20
E	0.64	0.54	1.0	0.7	0	0.58

Step 10:	A	B	C	D	E	T(I)
A	0	-.5	-1.	-1.37	-2.04	-.98
B	0.5	0	-.6	-.9	-1.57	-.51
C	1.0	0.6	0	-.4	-1.	0.04
D	1.37	0.9	0.4	0	-.7	0.39
E	2.04	1.57	1.0	0.7	0	1.06